Using Blinded Analytics to Monitor Data Quality in Psychiatry Programs with Centrally Rated Endpoints: Lessons Learned in Phase 2 and Implementation into Phase 3

Todd M. Solomon, PhD
Director | Clinical Development
Mind Medicine

Disclosures

Full time Employee and Shareholder - Mind Med

Former employee and current shareholder - Signant Health

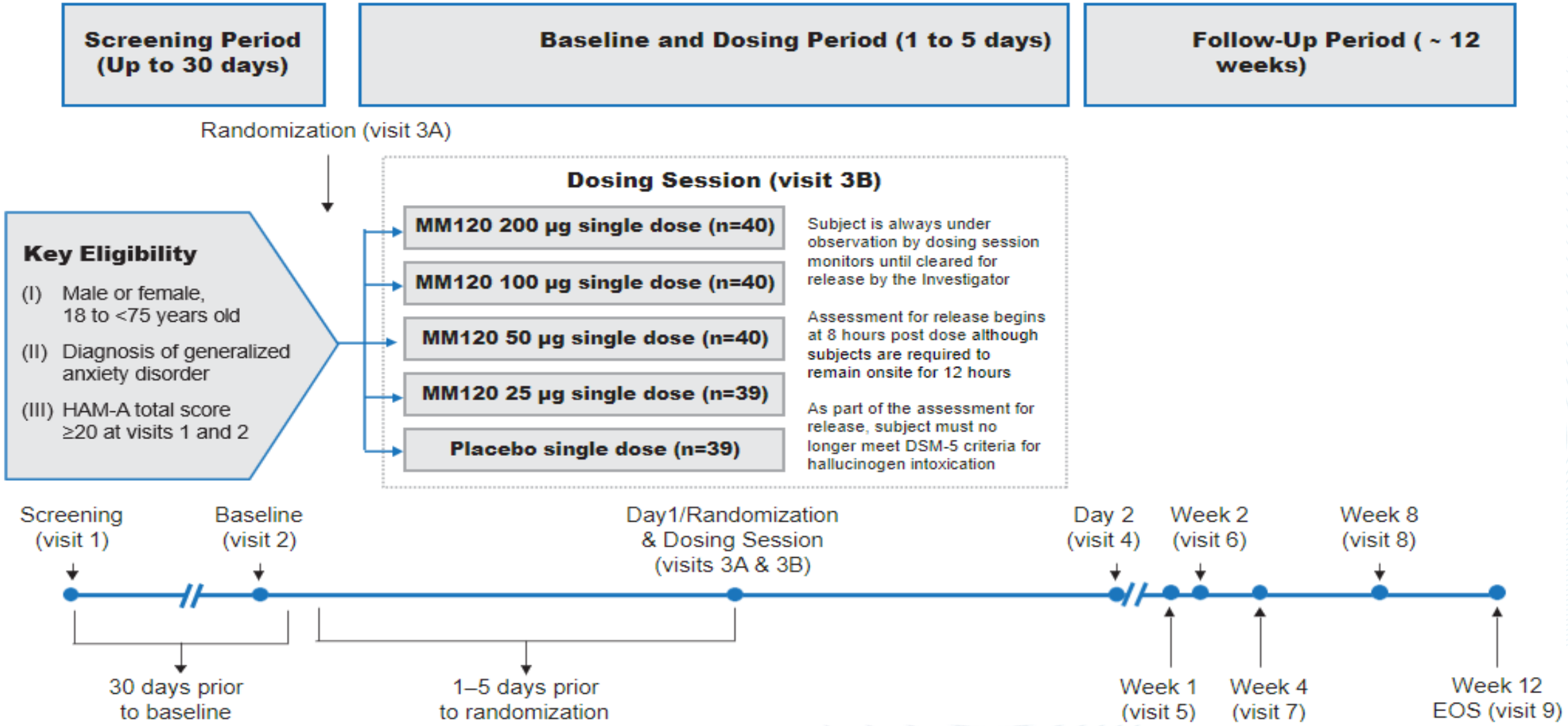
Outline

- “Psychedelic” Clinical Trials Considerations
- Blinded Analytics in a Phase 2 Program
- Considerations and Implementation into Phase 3 Pivotal Studies
- The Future of Data Oversite

”Psychedelic” Trial Design adds Methodological Complexity

- Drug to Site: Import, S1 License, Storage, Dispensing
- Single Dose Paradigm
- Dosing Day | Oversite, Setting, Personal
- Functional Unblinding
  - Central Raters (blinded to protocol, visit, etc.)
  - Measurement of blinding (participants, central raters, site raters)
  - Post Baseline primary endpoint blinding of sites
  - Firewall between dosing session monitoring and endpoint rating
  - ”Active PBO / Low Dose”

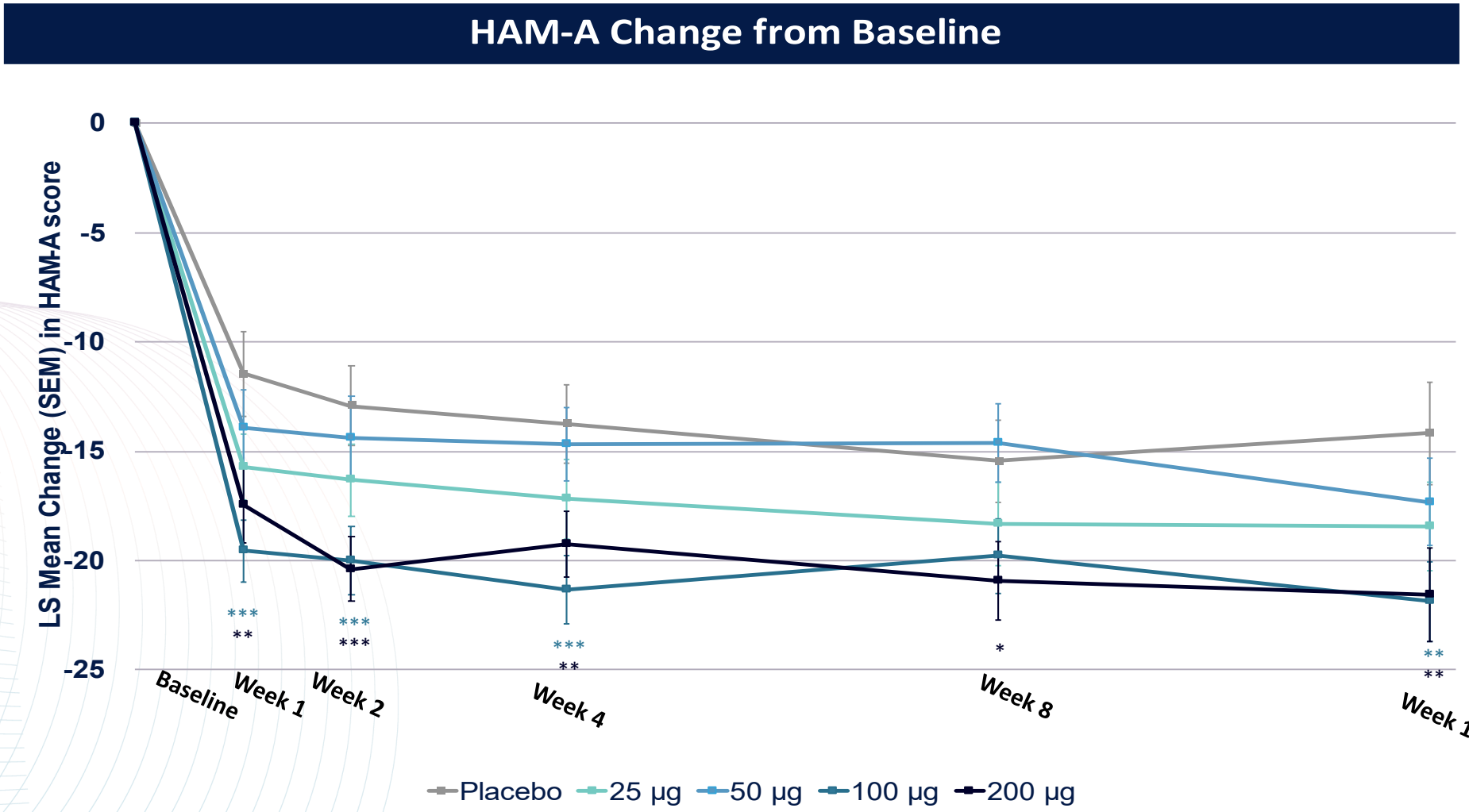
Phase 2b Trial Schematic[1]



1. Source: Study MMED008 internal study documents.

µg: microgram; HAM-A: Hamilton Anxiety Rating Scale;

Change in HAM-A Scores through Week 12 (FAS)[1]



MM120 100 µg
Change from Baseline[1]

- Week 4: -21.3 points
- Week 12: -21.9 points

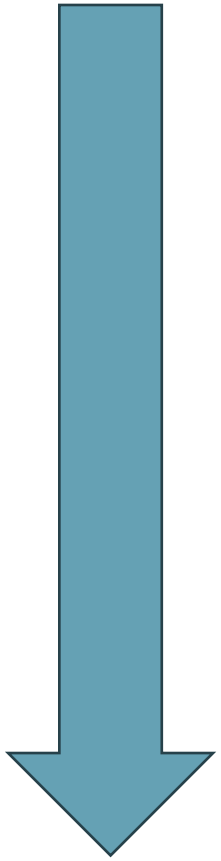
Improvement over Placebo[2]

- Week 4: -7.6 pts, p=0.0004
- Week 12: -7.7 pts, p=0.003

1. Source: Study MMED008 internal study documents and calculations. Full analysis set population.
µg: microgram; FAS: Full Analysis Set; HAM-A: Hamilton Anxiety Rating Scale; Statistical comparison vs. the placebo group using ANCOVA

Analytics & Data Oversight

Top Down

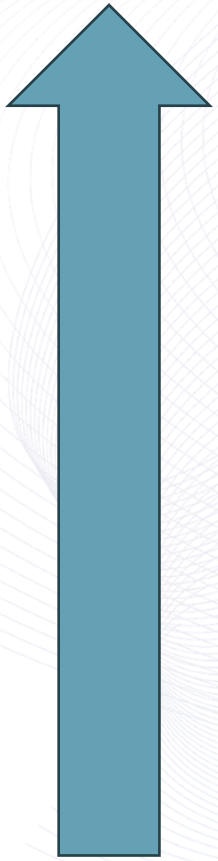


❖ Study Level Data in Aggregate
- Descriptive Statistics
  - Counts, Mean, Med, Mode, SD

❖ Site Level Data in Aggregate
- Site Compared to Study Means
- Inter Site Comparisons

❖ Study Level Rater Performance
- Meta Data (Time of interviews, Secondary Ratings, Discordance, inter/intra comparisons)

Bottom Up



❖ Individual Scale Scores

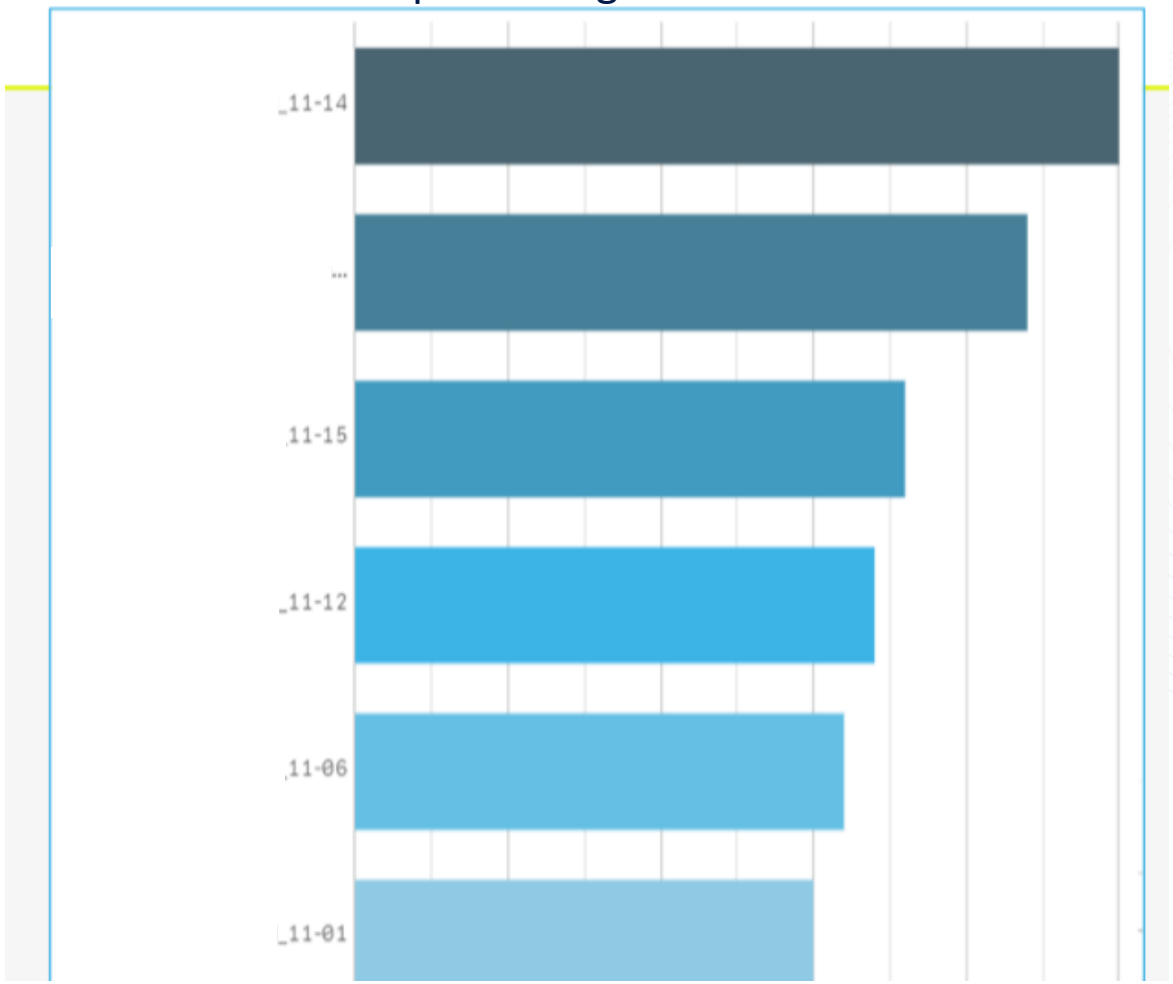
❖ Site Rater Performance
- Secondary Review
- Meta Data

❖ Central Rater Performance
- Training and Certification
- In Study Performance
- ICC

❖ Visit Level Discordance
- Alignment between scales measuring similar constructs

Study Level Analytics P2

Top Enrolling Site



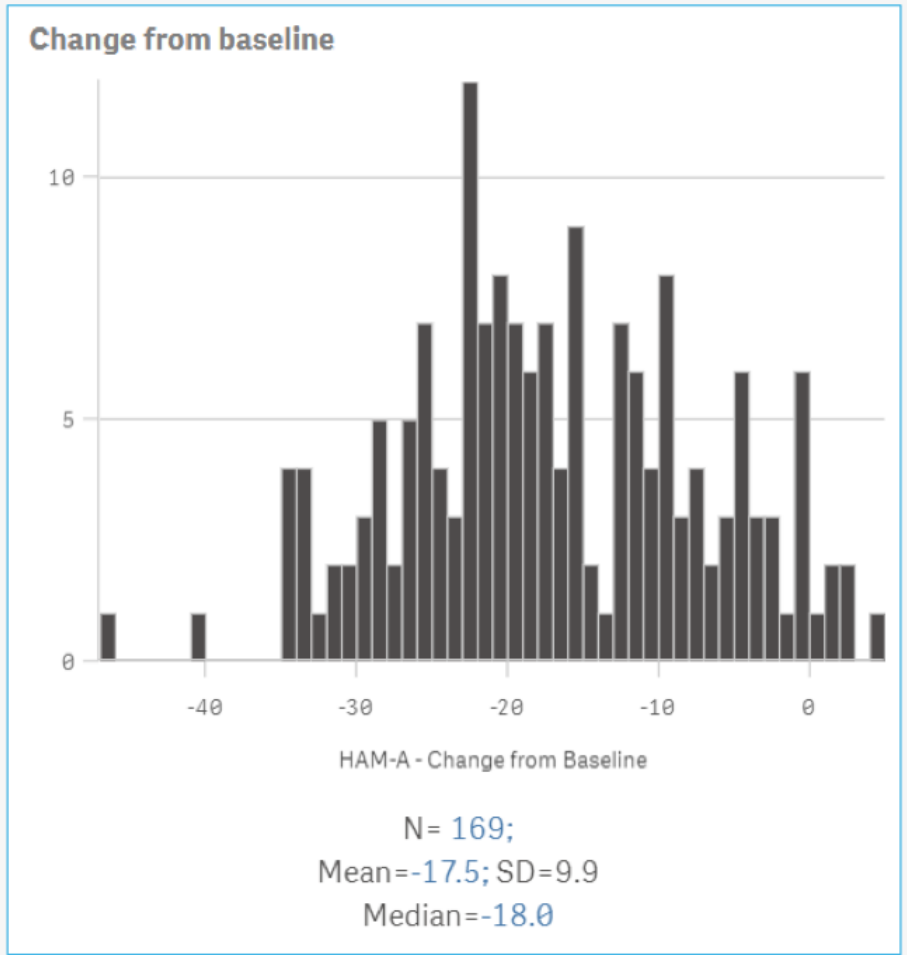
	Since FPI	New Data
Visit	Count of visits	Count of New visits
Totals	**1931**	**90**
Visit 1 - Screening	481	0
Visit 2 - Baseline	234	0
Visit 3A - Day 1	198	0
Visit 4 - Day 2	197	0
Visit 5 - Week 1	193	1
Visit 6 - Week 2	182	13
Visit 7 - Week 4	171	25
Visit 8 - Week 8	137	28
Visit 9 - Week 12/Early Term	138	23

lew data date range: 9-Sep-2023 through 13-Oct-2023

Study Level Analytics P2

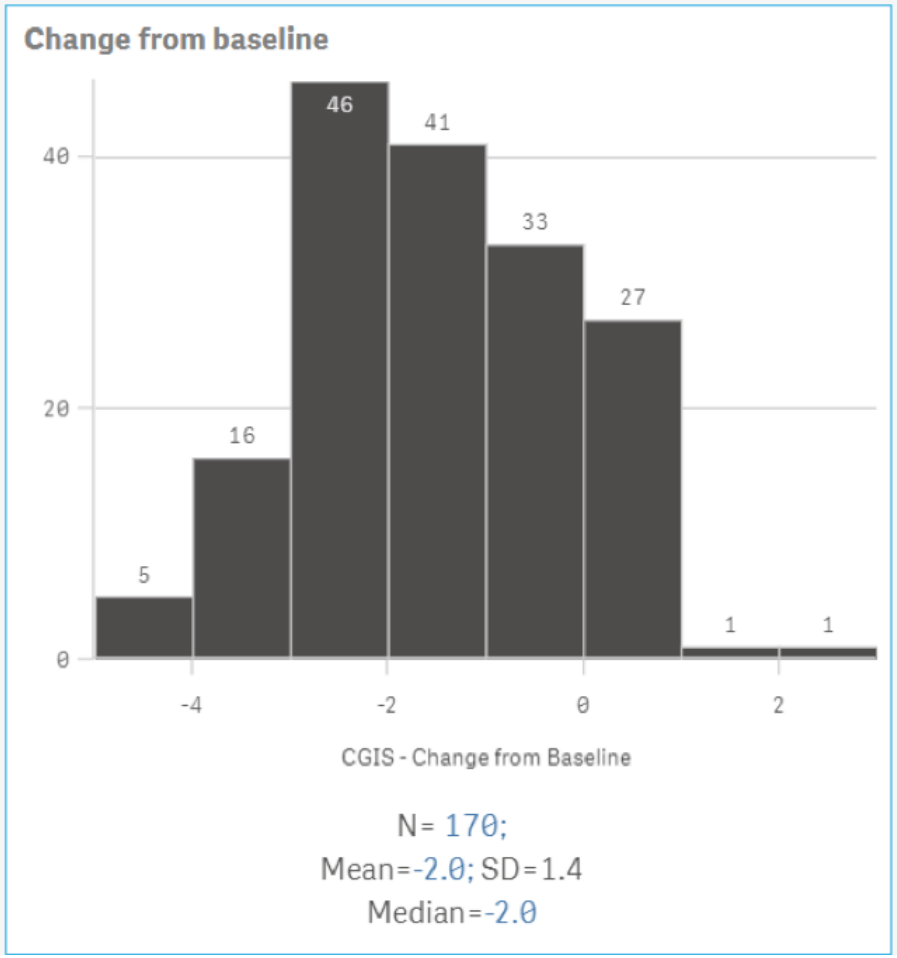
HAM-A and CGI-S Week 4/Visit 7 Change Score Distributions

HAM-A Week 4/Visit 7 Change Distribution



N= 169;
Mean=-17.5; SD=9.9
Median=-18.0

CGI-S Week 4/Visit 7 Change Distribution



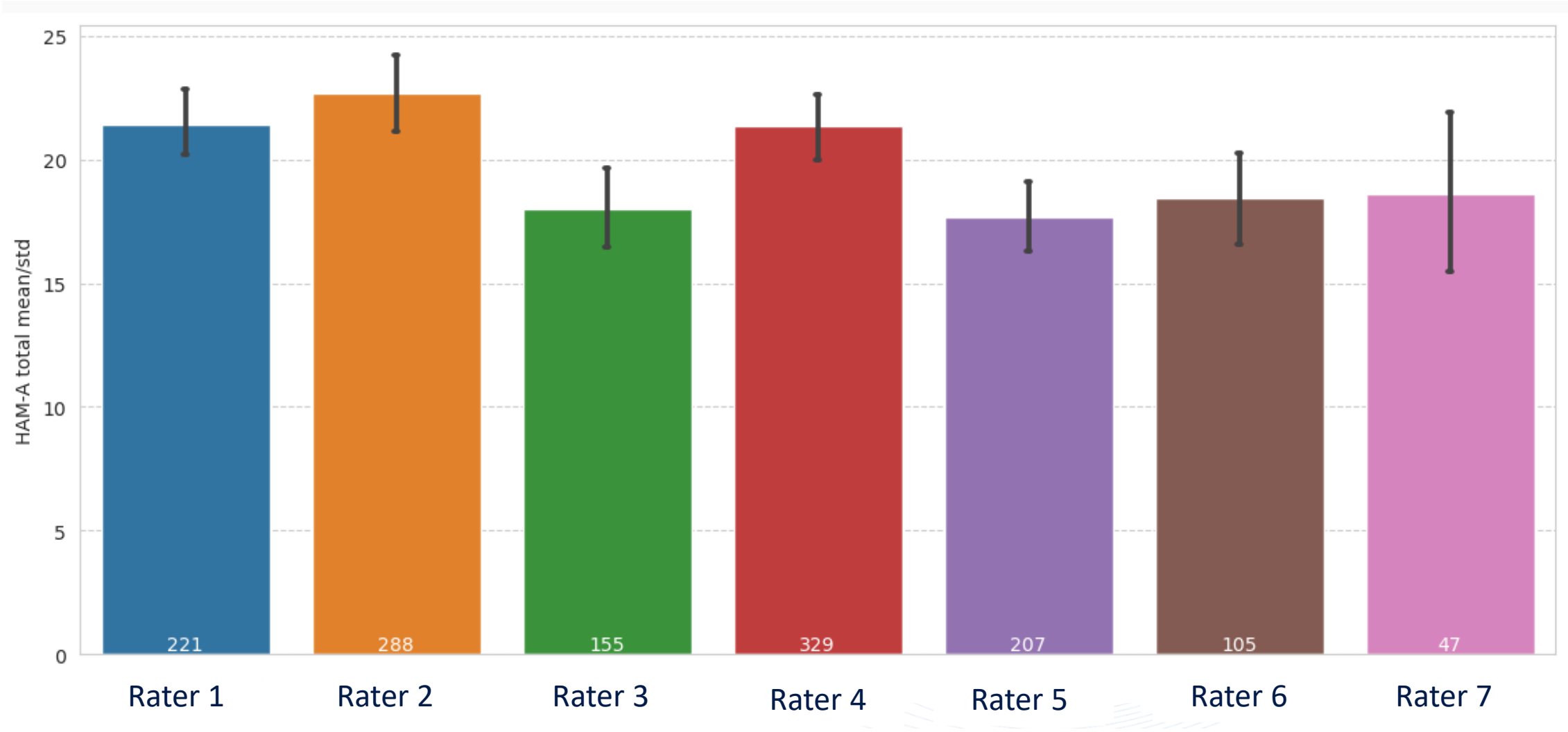
N= 170;
Mean=-2.0; SD=1.4
Median=-2.0

The trouble with “Central Rating”

- To help mitigate functional unblinding, our phase 2 program employed the use of central raters who were blinded to participant, visit, protocol, etc.
  - Smaller group of raters collecting primary / key secondary endpoints
  - Outsized influence on study data
  - Management and oversite provided by 3rd party
- Central Rater Oversite
  - Rater Training and Certification (RTC)
  - In study performance methodology
  - Interclass Correlation Coefficients (ICC)

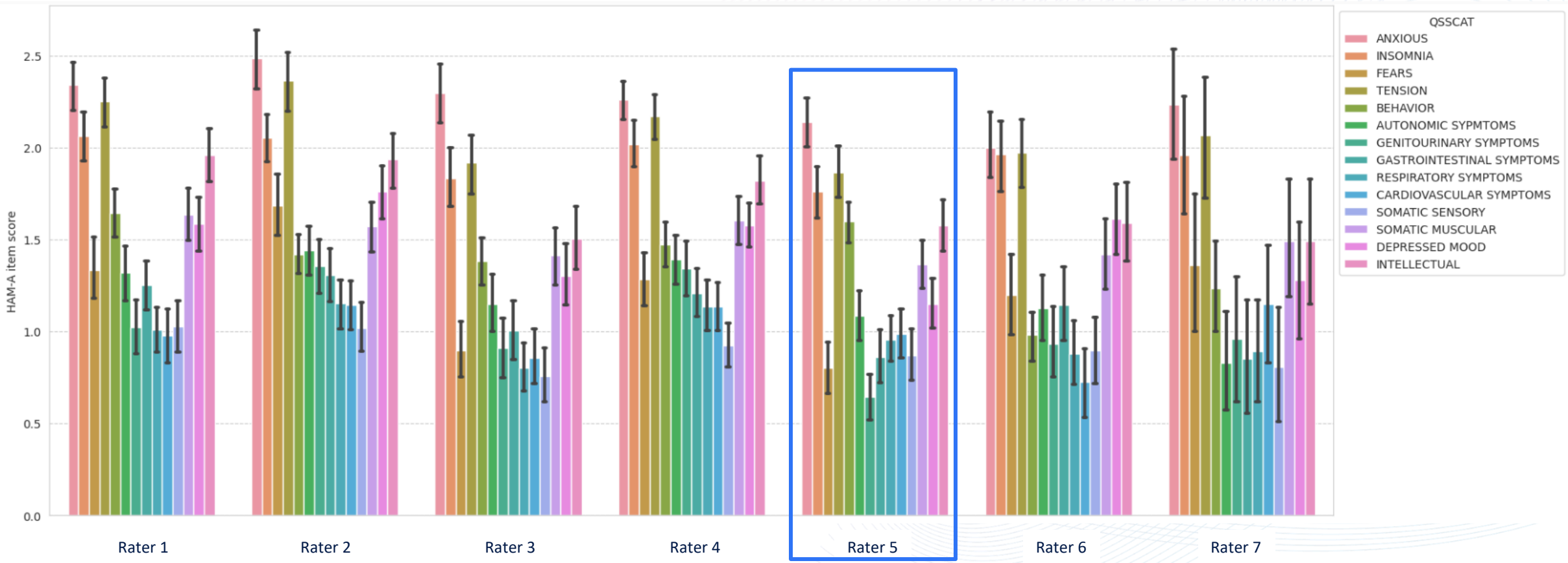
Central Rater Analytics P2

HAM-A Distribution x Rater



Central Rater Analytics P2

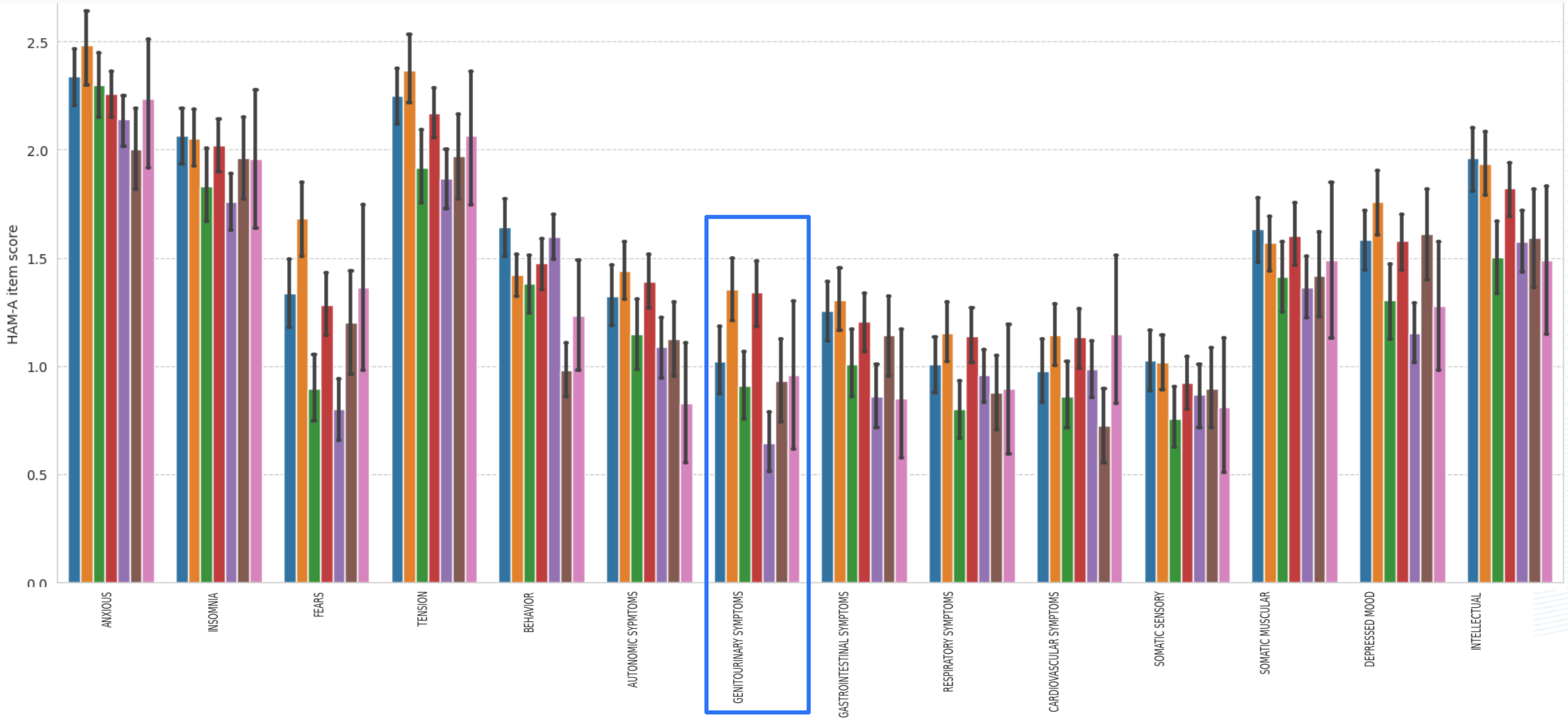
HAM-A Item Score x Rater (intra)



Central Rater Analytics P2

HAM-A Item Score x Rater (inter)

- One rater was consistently scoring lower across 65% of items



Central Rater Analytics P2

rater / item	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	Rater 6	Rater 7
ANXIOUS	0.414295	0.000484	0.883792	0.612566	0.037358	0.005135	0.744785
AUTONOMIC SYPMTOMS	0.502874	0.004522	0.133730	0.033626	0.008469	0.148844	0.005413
BEHAVIOR	0.001131	0.523643	0.322517	0.626671	0.015746	0.000000	0.110646
CARDIOVASCULAR SYMPTOMS	0.501769	0.047262	0.052976	0.045018	0.593082	0.004771	0.445136
DEPRESSED MOOD	0.316640	0.000038	0.016344	0.240897	0.000001	0.370331	0.153751
FEARS	0.316619	0.000000	0.000237	0.657490	0.000000	0.648927	0.565412
GASTROINTESTINAL SYMPTOMS	0.113839	0.006238	0.117218	0.238946	0.000113	0.993718	0.076160
GENITOURINARY SYMPTOMS	0.370053	0.000036	0.051206	0.000020	0.000000	0.172010	0.449787
INSOMNIA	0.139799	0.123965	0.103252	0.306272	0.002726	0.973752	0.959875
INTELLECTUAL	0.007099	0.005798	0.003403	0.317085	0.012112	0.116433	0.104183
RESPIRATORY SYMPTOMS	0.813083	0.019894	0.004854	0.026323	0.314669	0.133996	0.387269
SOMATIC MUSCULAR	0.112783	0.436728	0.192950	0.147676	0.024948	0.321457	0.831645
SOMATIC SENSORY	0.132399	0.109962	0.035812	0.957640	0.409546	0.756391	0.446716
TENSION	0.090844	0.000086	0.013893	0.459106	0.000304	0.142000	0.690189

Parametric Analysis

[83]

```
df_ps
```

item	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	Rater 6	Rater 7
ANXIOUS	0.541070	0.000000	0.991895	0.339690	0.001328	0.009955	0.997564
AUTONOMIC SYPMTOMS	0.404182	0.004850	0.166230	0.007369	0.132914	0.090296	0.099854
BEHAVIOR	0.010839	0.303816	0.480047	0.005246	0.000001	0.000000	0.109180
CARDIOVASCULAR SYMPTOMS	0.856465	0.460234	0.182113	0.035441	0.040459	0.086860	0.941158
DEPRESSED MOOD	0.650650	0.000000	0.171163	0.903652	0.000285	0.429919	0.627138
FEARS	0.055533	0.000000	0.000404	0.917558	0.000013	0.995866	0.980144
GASTROINTESTINAL SYMPTOMS	0.003542	0.130331	0.198887	0.126616	0.000060	0.721715	0.205555
GENITOURINARY SYMPTOMS	0.936405	0.000029	0.068521	0.000106	0.000000	0.157718	0.998129
INSOMNIA	0.072864	0.008366	0.213155	0.977138	0.000328	0.999810	0.995151
INTELLECTUAL	0.078912	0.000010	0.000701	0.998958	0.001193	0.399030	0.414630
RESPIRATORY SYMPTOMS	0.968501	0.149058	0.109913	0.021873	0.365360	0.424253	0.878536
SOMATIC MUSCULAR	0.374068	0.848183	0.087045	0.005334	0.013625	0.268473	0.995002
SOMATIC SENSORY	0.414198	0.324462	0.238659	0.214050	0.977615	0.999825	0.908703
TENSION	0.106401	0.000000	0.040539	0.999126	0.000002	0.609832	1.000000

Non-Parametric Analysis

[84]

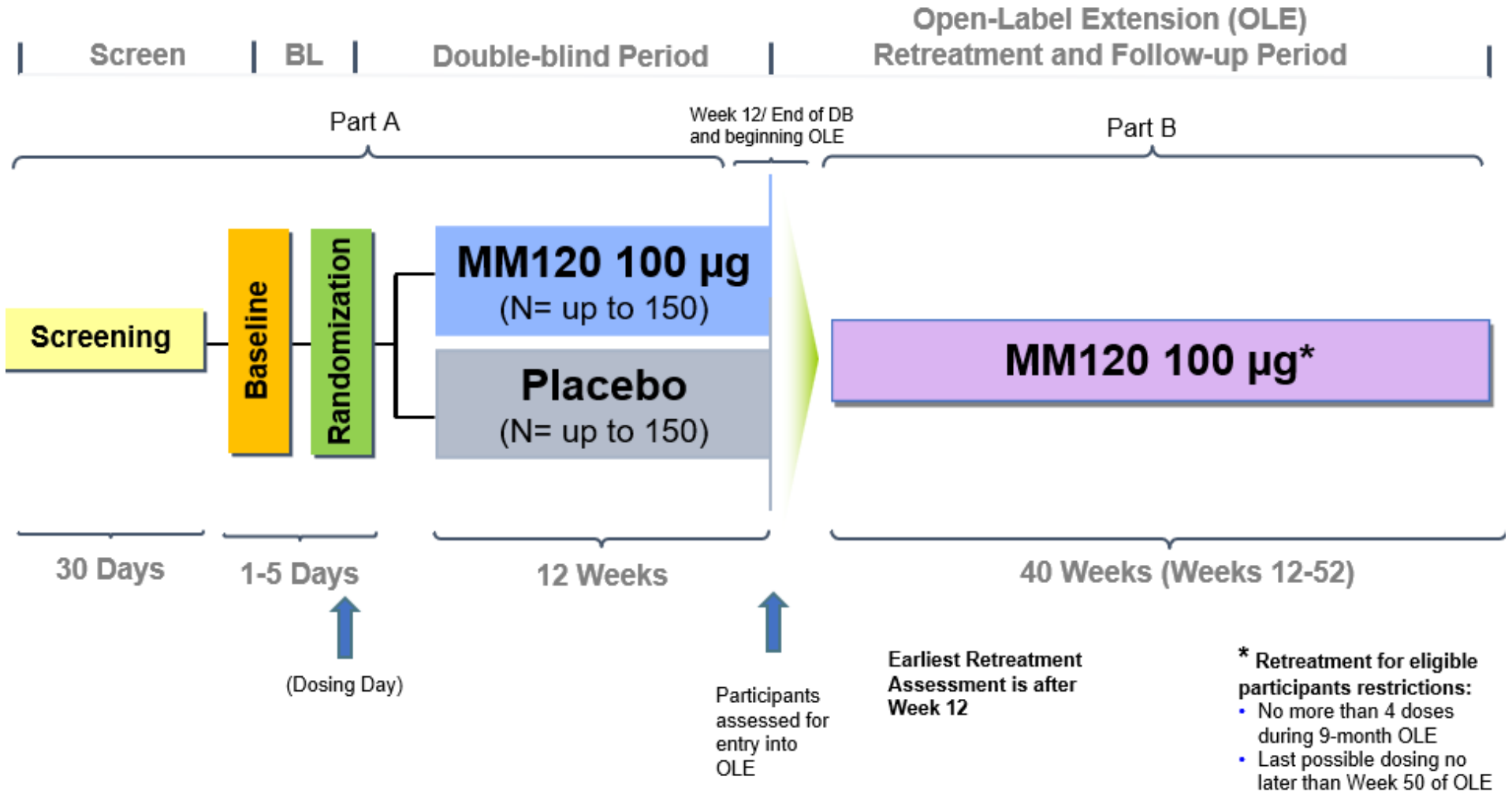
MM120-300 Study Design

Voyage

US only

SELECT ENTRY CRITERIA

- Men and Women
- Ages 18-74
- Diagnosis of GAD
- HAM-A ≥ 20
- MADRS Items 1, 7, and 8 ≤ 2



BL = Blinded; DB = Double-blind; HAM-A = Hamilton Anxiety Rating Scale; N = number of subjects; OLE= Open Label Extension

15

Blinded Sample Size Recalculation

- Both phase 3 programs have an adaptive component built into their design that requires active monitoring of data

- As referenced in the Adaptive Designs Based on Non-Comparative Data section of the 2019 FDA Guidance for Industry titled, “Adaptive Designs for Clinical Trials of Drugs and Biologics” (Gould 1992)

- Variables:
  - Placebo adjusted difference for HAM-A (P2)
  - Model based estimate of Standard Deviation (P2)
  - Effect Size (P2)
  - Power (P3)
  - Standard Deviation (P3 observed)
  - % of participants who ET

Designing Analytics and Oversite for Pivotal Studies

Goal: Design a robust program for monitoring study data and rater performance across multiple Phase 3 programs:

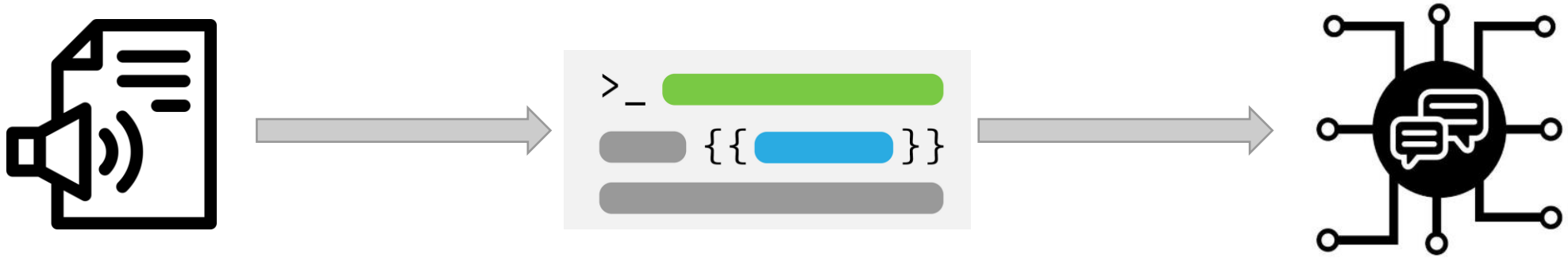
- eCOA – Platform based solution to allow for real time data collection and oversite
  - Web based, device agnostic (Central Ratings, Site Ratings, ePRO App)
  - Alerts
  - Notifications
  - Real time Intra-visit algorithms
- Study Analytics
  - Risk based algorithmic monitoring of at Site and Study level data
  - GUI for data visualizations and interactions
- Central Rater Monitoring
  - Robust RTC
  - Continuous Inter and Intra analysis of central rater performance to ensure individual raters don't drift or skew data
- Hammy

eCOA – Electronic Clinical Outcome Assessment
GUI – Graphical User Interface
RTC – Rater Training and Certification

Could Large Language Models Help?

- Large Language Models (LLMs) excel in understanding text and analyzing large data sets
- Open source LLMs allow us to host the model locally, ensuring data security
- LLMs can be trained for specific tasks
- We can train an open-source LLM to score HAM-A interviews and generate its own score as a quality check on central raters, providing a comprehensive and consistent method for rater oversight

How to Train Your Model

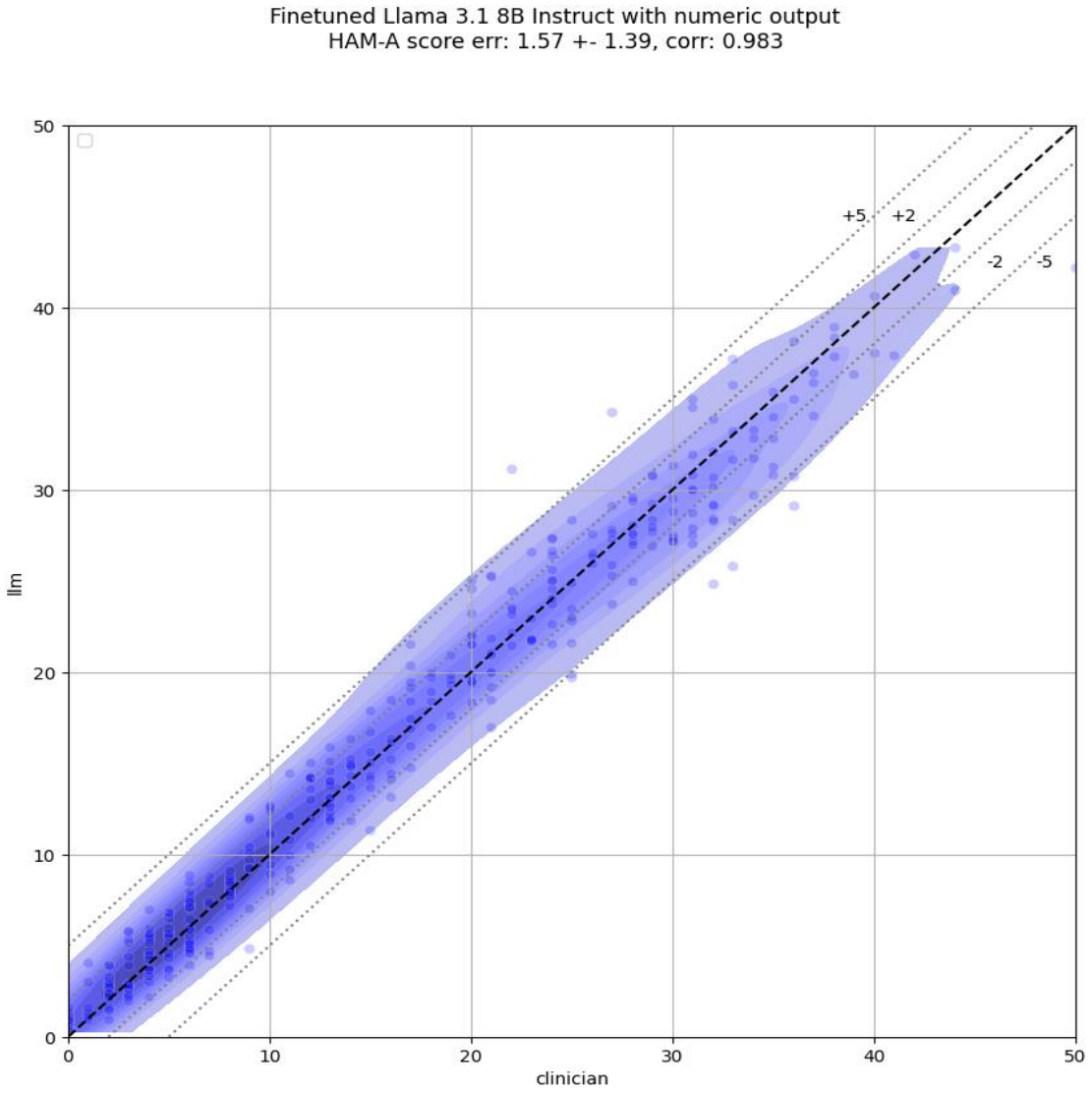


- 1500 sessions of HAM-A audio from Ph 2
- 21k individual symptom ratings
- Audio transcribed to text-based dataset

- Design prompt for LLM
- Prompt on symptom level basis
- Each prompt contains symptom utterance and rater's guidelines

- Finetune open source LLM
- Prediction of each symptom score

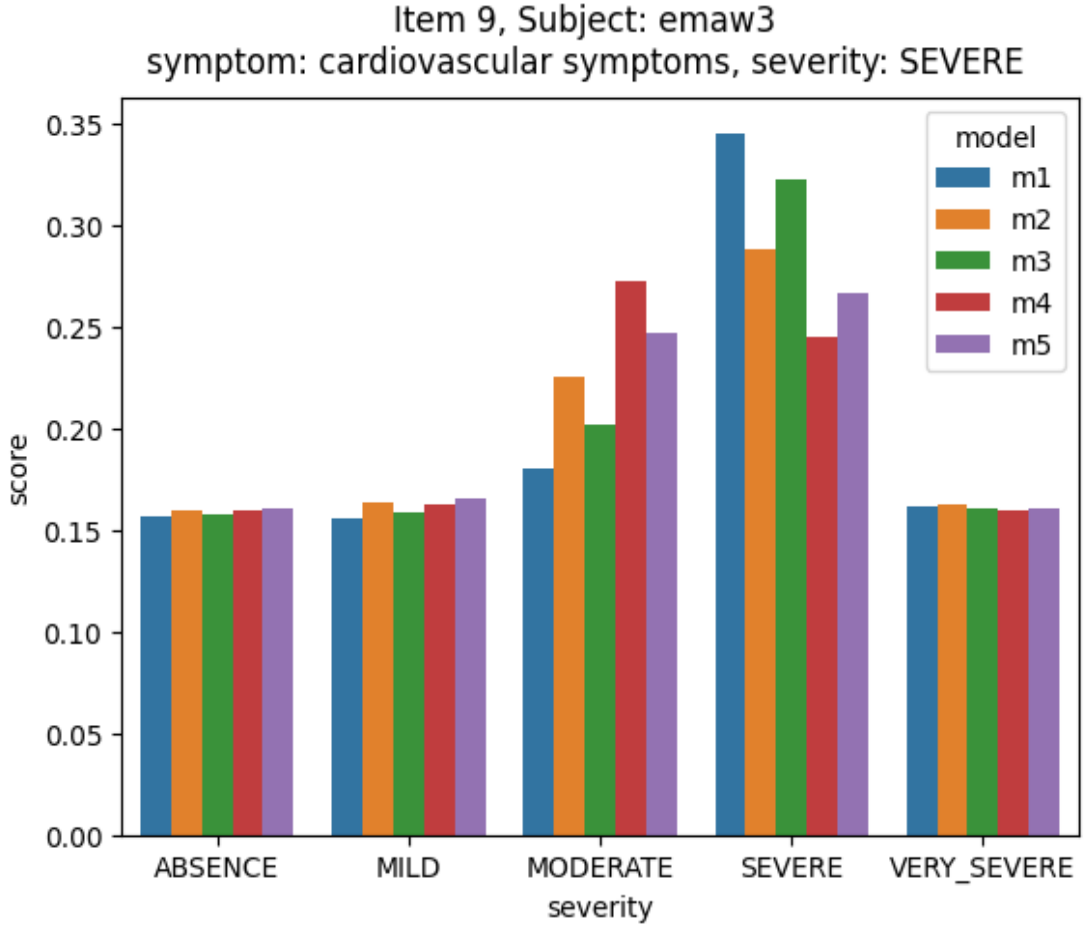
Testing HAMMY vs Ph 2 data



- Each dot represents one HAM-A assessment from Phase 2
  - X-axis is the clinician score
  - Y-axis is the HAMMY score
- Regression line represents exact agreement between the clinician and HAMMY
- Average difference between HAMMY total scores and clinician total scores is 1.57 (+/- 1.39)
  - Pearson Correlation = 0.98
  - Outliers point to problematic clinical ratings

Application to Phase 3

- Model works well on the Phase 2 data, but we continued to refine the model before applying to Phase 3
- We split the model into 5 instances, or sub-models
- These 5 sub-models act as a "committee", each voting on the score they think is most appropriate
- This allows for:
  - Increased Robustness: If one sub-model has an outlier rating, it will be canceled out by the others
  - Approximating Confidence: Items where the sub-models disagree on scoring indicate that particular item was difficult to assess.



Mock interview

High severity case

	EMA	HAMMY
1. anxious mood	3	3
2. tension	3	3
3. fears	3	3
4. insomnia	3	3
5. intellectual symptoms	3	3
6. depressed mood	3	3
7. muscular symptoms	3	3

	EMA	HAMMY
8. sensory symptoms	3	3
9. cardio	2	2
10. respiratory	2	2
11. gastrointestinal	2	2
12. genitourinary	2	2
13. autonomic symptoms	2-4	3
14. patient behavior	2	2

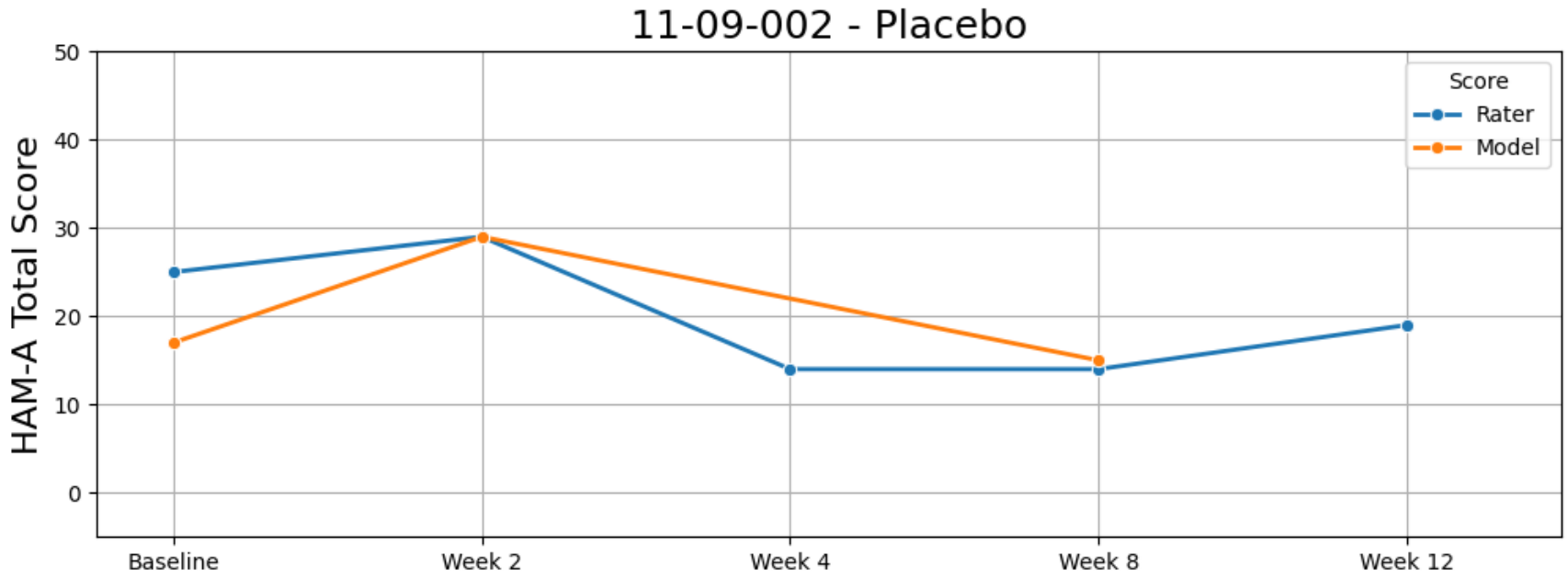
EMA overall score is **36 - 38**
HAMMY overall score is **37, 100% correct**

Visualizing a single interview

	Anxious Mood	Tension	Fears	Insomnia	Intellectual Symptoms	Depressed Mood	Somatic Muscular Symptoms	Somatic Sensory Symptoms	Cardiovascular Symptoms	Respiratory Symptoms	GI Symptoms	Genitourinary Symptoms	Autonomic Symptoms	Behavior	Total
Rater	3	4	0	3	3	0	0	0	0	0	2	4	3	3	25
Model	3	2	0	3	2	0	0	0	0	0	2	0	3	2	17

Participant could've been excluded (Baseline visit)

Ended up being in Placebo with a significant response (-11 pts, -44%) at Week 8



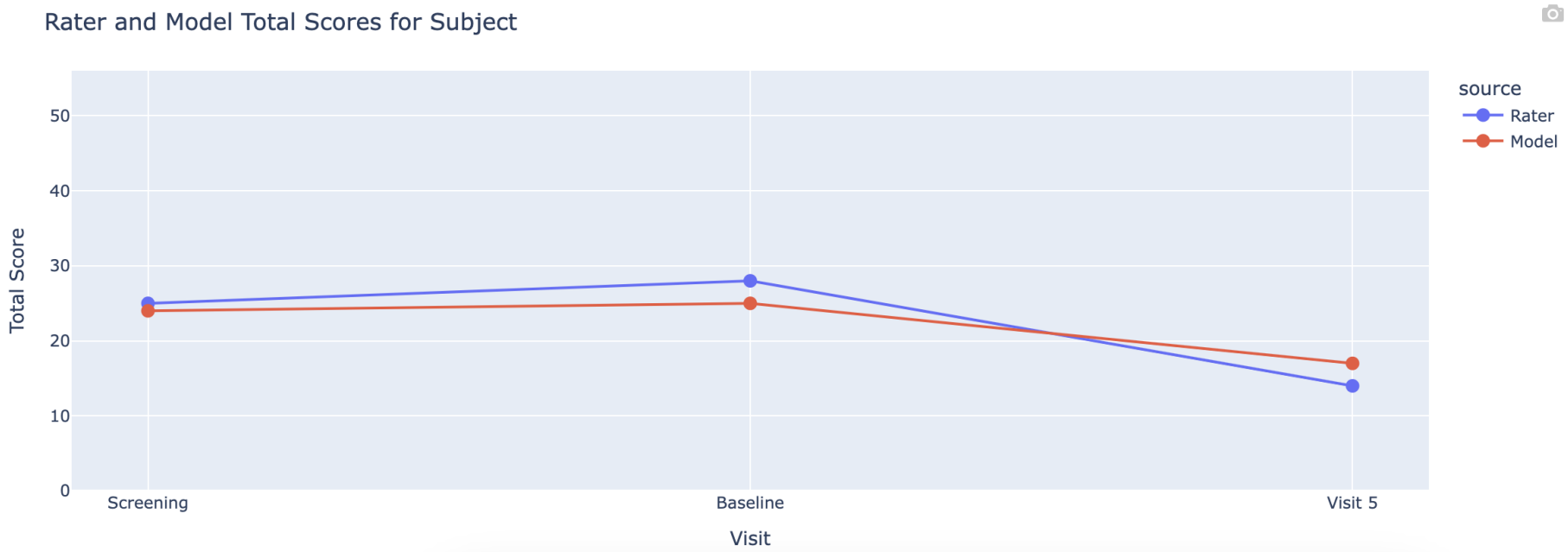
Integration into P3 Oversite – Study Dashboard

Details for Voyage - - Visit 5

Source	Audio	Tra...	Items	#0 OP	#1 AM	#2 TN	#3 FR	#4 IN	#5 IS	#6 DM	#7 SM	#8 SS	#9 CV	#10 RS	#11 GI	#12 GU	#13 AU	#14 PB	#15 CL
Rater					2	2	0	2	2	0	2	0	0	0	3	0	0	1	
Model					2	2	0	2	2	0	2	2	0	0	3	1	0	1	
Confide...					High	High	High	High	High	High	High	Mid	High	Mid	High	Mid	High	High	
Duration	47:55	43:59	37:47	00:57	03:46	04:03	01:28	03:39	03:16	01:20	03:34	05:46	00:28	01:18	03:29	02:04	01:03	01...	00:10

Show Transcript | Hide Subject History | Show Confidence

Rater Total	14
Model Total	17
Model Min	16
Model Max	18
Abs Diff	3
Perc Diff	21.4%
Accum Diffs	3.0
Num Diffs	2.0



Thank You